Block Estimating of Spatial Yield Data and its Uncertainty

SHUFENG HAN   HanShufeng@JohnDeere.com
Deere & Company, Urbandale, IA 50131, USA

SALLY M. SCHNEIDER
USDA-ARS, Fresno, CA 93723, USA

ROBERT G. EVANS
USDA-ARS, Sidney, MT 59270, USA

JOAN R. DAVENPORT
*Irrigated Agricultural Research and Extension Center, Washington State University,
Prosser, WA 99350, USA*

Abstract.   On-the-go yield monitors have been available for both grain and bulk crops. Most of the yield monitors today provide yield measurement at a fixed time interval. Conversion of these point yield data into raster yield maps for further analysis is necessary. In this study, a data-blocking procedure is proposed to create raster yield maps from point yield data. The blocking procedure includes: (1) converting the fixed-time-interval data into fixed-distance-interval data; (2) using a moving average algorithm to estimate a cell value when there are sufficient data points within the cell; (3) using a geostatistical algorithm to estimate a cell value when there are not enough data points within the cell but values of its neighboring cells are known; and (4) calculating an uncertainty index for each cell value estimation. An example application of the yield-blocking procedure with potato harvest data in 1996 was given.

Keywords: crop management, yield, mapping

Introduction

Yield mapping is one of the most successful technologies currently used in site-specific crop management. Knowledge of within-field yield variation can help identify site-specific management needs of each individual field. On-the-go yield mapping systems now have been available for many different crops (Borgelt and Sudduth, 1992; Schneider *et al.*, 1996; Walter *et al.*, 1996). A typical yield mapping system includes a yield monitor and positioning equipment. Yield monitors for grain crops commonly use sensors to measure the grain flow rate and moisture at the exit of the clean grain elevator. For bulk crops such as potatoes, load cells are placed under the harvester conveyor system to determine the crop weight. Currently, the most popular positioning equipment is the Global Positioning System (GPS) receiver.

The errors of a yield mapping system come from many different sources (Blackmore and Marshall, 1996). Two of the most significant sources of errors have been identified as the varying harvest width and the time delay between actual harvest position and

sensing of the crop yield. Various techniques have been developed to reduce these errors. For examples, Vansichen and Baerdemaeker (1991), and Reitz and Kutzbach (1996) used an ultrasonic distance transducer to measure the unused portion of the combine header width for determining the width of swath cut. An accuracy of better than 0.02 m was reported. As a software alternative, Han *et al*. (1997) developed a bitmap method for determining effective combine cut width. If a high-accuracy positioning system is available, the bitmap method can be used to accurately calculate the actual harvest area for each instant yield measurement. A similar concept called potential mapping was proposed by Blackmore and Marshall (1996) to reduce the errors associated with the use of a fixed harvest width, although the question as how to calculate the actual harvest area from the positional information was not addressed in the report. The calculation of time delay seems to be more complicated, since the crop transport dynamics must be considered. Searcy *et al*. (1989) used a first-order time delay equation to describe the combine grain transport dynamics. However, the parameters of the transfer function were difficult to determine. At present, a fixed time delay is often applied for the whole field or for each individual path.

After the errors associated with the harvest width and time delay have been corrected, a reasonably accurate set of point yield data can be obtained. The point yield data are discrete in space, with each data point representing an average yield over an area surrounding that point. The distribution of these points in a field is irregular due to changes in harvester speed, removal of obviously erroneous data, missing of data points, and other factors. For the purposes of visualization and analysis, it is often necessary to interpolate the irregularly spaced point yield data to a regular grid. We call this process data blocking.

Two groups of interpolation methods have been commonly used for blocking yield data: kriging and inverse distance. The kriging procedure can provide the best linear unbiased estimates and has been used extensively to evaluate variability of soil physical and chemical properties (Burgess and Webster, 1980a,b; Meirvenne and Hofman, 1989; Paz *et al*., 1996). The kriging procedure is most useful when the available data set is sparse and the spatial correlation of the property is clear. However, point yield data are often very dense, which makes the kriging procedure very inefficient in calculation. Although many commercial yield-mapping software programs use the inverse distance method instead of kriging, the reasons for using this method are often unclear. Moreover, many yield-mapping software programs only provide the yield maps as the end product, without indicating the accuracy of these maps. Clearly, a yield map created by a sparse data set is most likely not as accurate as the one created by a dense data set. The objective of this research is to develop an efficient yield-blocking procedure for yield map generation and for the accuracy assessment of the generated yield map.

Materials and methods

The proposed yield-blocking procedure (Figure 1) includes: (1) converting the point yield data set (often in fixed time interval) into a fixed-distance-interval data set; (2) using a moving average algorithm to estimate a cell value when there are sufficient
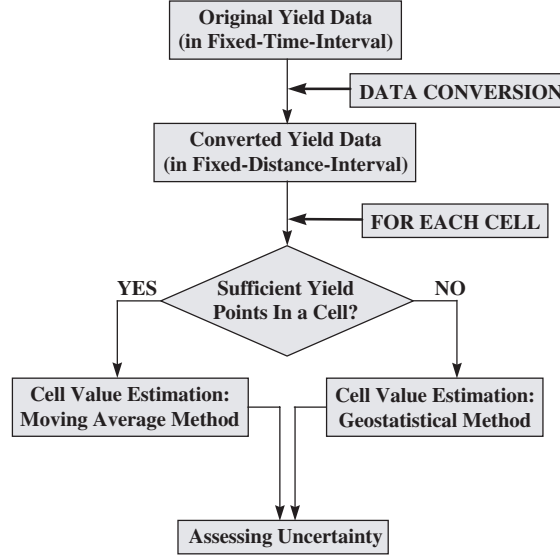
Figure 1.   Flowchart of the yield blocking procedure.

data points within the cell; (3) using a geostatistical algorithm to estimate a cell value when there are not enough data points within the cell but values of its neighboring cells are known; and (4) assessing the uncertainty for each cell value estimation.

Data conversion

Most yield mapping systems collect yield data at a fixed time interval. Since the travel speed of the harvester may vary in the field, the distance between two yield measurement points is usually not fixed. Suppose that we have n yield measurements, y_i $(i = 1, 2, \ldots, n)$, in an area A, and that each measurement y_i is taken from a subarea A_i. The average yield, $\bar{y}$, in the area A is:

$$\bar{y} = \sum_{i=1}^{n}(y_i * A_i) \Big/ \sum_{i=1}^{n} A_i \tag{1}$$

The arithmetic average of the n yield measurements, $\bar{y}'$, is:

$$\bar{y}' = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2}$$

Unless all subareas A_i are of equal size, the average yield in the area A does not equal the arithmetic average of the n yield measurements. As an example, suppose we have two yield measurement points, $y_1 = 10$ Mg/ha and $y_2 = 15$ Mg/ha. These two measurements are taken from two unequal subareas, $A_1 = 20$ m^2 and $A_2 = 25$ m^2, respectively. The average yield for the entire area A ($A = A_1 + A_2$) is 12.8 Mg/ha. However, the simple arithmetic average of these two yield measurements is 12.5 Mg/ha. The *data conversion* procedure converts the original yield data set into a new yield data

set so that each yield point in the new data set corresponds to the same size of area. This new yield data set will be called the converted yield data set.

Let y_i be a point yield taken from a subarea A_i. A_i is calculated by:

$$A_i = W_i * L_i \tag{3}$$

where W_i is the effective harvest width and L_i is the harvest distance for the ith yield measurement. Although recorded at discrete points, yield is a continuous variable. If A_i is small enough, we can assume that yield is a constant y_i within each subarea A_i. Practically, we can take k_i yield samples along the harvest distance L_i to represent the continuous distribution of the yield in the subarea A_i. The k_i yield samples have the same value y_i and are evenly spaced with a small separation distance d along the harvest distance L_i. Thus the converted yield data set includes a total of $m = \sum_{i=1}^{n} k_i$ yield points. Assuming a constant effective harvest width, then, each yield point in the converted data set corresponds to the same size of area, $W_i * d$.

The converted yield data set is based on a fixed-distance interval, d. Unlike the original yield measurements, each yield point in the converted yield data set has the same weight. This allows many conventional statistical formulas to be easily applied to the converted yield data set. For example, the average yield in an area A is the arithmetic average of all the yield points within that area.

Another feature of the data conversion is that the method provides yield estimates within the gaps between any two consecutive yield measurement points. These gaps are artifacts due to the time interval selected to record yield data. Treating these gaps as areas with no yield does not make sense. With the data conversion algorithm, the gaps can be reduced small enough (by using a small fixed-distance interval d) so that the converted yield data set can approximate the continuous yield distribution. However, some gaps such as those caused by momentary GPS signal loss should still be treated as areas of missing yield measurements.

Cell value estimation using a moving average algorithm

The fixed-distance interval data set is not a grid because the selected fixed-distance interval d is usually much smaller than the harvest width W. The next step of the data blocking procedure is to convert this data set into a regular grid. The cell size selection for the latter grid is primarily based on the minimum management size for the field.

With normal operation of yield mapping systems, yield data are often collected over the entire field. For most cells, there will be a large number of yield data points taken within the individual cell. The average value of the measured yield in a cell would best represent the cell value. When the converted yield data set is used, the average yield within a cell is simply the arithmetic average of all the yield points within that cell. This process estimates cell values for most of the cells.

Cell value estimation using a geostatistical algorithm

For some cells in the field, there may not be sufficient area coverage by the measured yield data within the cell. This may be due to a number of factors such as removal of

obviously erroneous data, or missing data points. When the area coverage by the measured data is not large enough, using an average measurement yield for the cell value is not reliable. When no measurement data are available within a cell, the cell value must be estimated from data outside the cell.

Although an inverse distance weighting method is often used for yield data interpolation and its accuracy is comparable with the kriging method, the latter is preferred in our study because of its capability to establish a confidence level for each cell estimation. The theoretical background of kriging and the equations defining the kriging algorithm can be found in many references (Isaaks and Srivastava, 1989; Journel and Huijbregts, 1978; Matheron, 1963). A computer program developed by Bogaert *et al*. (1995) was used for the kriging analysis in this study.

In applying the kriging method to estimate a cell value, the main question is whether we should use the sample yields (from the converted data set) or the known cell values (from previous *moving average* estimation) as neighbors of the cell to be estimated. The main concern in using the first approach is that the number of yield points within a reasonable search distance around the estimated cell will be so large that the time required to solve the kriging system of equations is impractical. However, using the known cell-value approach is also not without question, since the known cell values are not original and they are not point values as required by kriging. However, if the cell size is small enough so that the known cell values can be treated as point data, the latter approach is a better choice.

Assessing the uncertainty for each cell estimation

The single cell value obtained from the previous steps is only a reasonable approximation of the true cell average. Qualitatively, we know that there is a difference between the estimated cell value and the true cell value. There is uncertainty associated with this difference. Other terms, such as reliability, confidence, accuracy, can also be used to express the same concept.

To assess the uncertainty, several factors that influence the estimation error should be considered. One obvious factor is the number of measured yield points (or the area coverage by the measured yield data within the cell). Assuming yield measurement error is negligible, increasing the number of measurements can generally reduce the estimation error. The second factor, which perhaps has more influence on the estimation error, is the nature of the yield distribution. Estimates will be more reliable for a well-behaved yield variable (such as one with a uniform distribution) than for a very erratic yield variable. Other factors, such as the spatial arrangement of the available yield data can also affect the accuracy of yield estimation.

Confidence intervals are the most familiar way of reporting uncertainty. A more general method is to use a probability distribution to assess the uncertainty. However, neither method provides a single index of uncertainty, which makes the graphical representation of uncertainty difficult. We propose using a single uncertainty index for each cell estimation.

For those cell values estimated by the moving average algorithm, the uncertainty index is defined as the percentage of the total area not covered by the measured yield

points. Mathematically, it is expressed as: $100 * (A_j - A'_j)/A_j$, where A_j is the total area of cell j and A'_j is the cumulative yield measurement area within cell j. This index is simple to calculate, and it captures the essential source of uncertainty: not enough measurements. Since the statistical properties of the yield distribution are not considered, the method may exaggerate the uncertainty in those cells with low percent coverage of the measured yield. For example, if the yield has a fairly uniform distribution in a cell, a few measurement points can provide a good cell-value estimate. However, the uncertainty index would still be calculated as very high. Fortunately, this situation is avoided in our procedure since we do not use the moving average algorithm when the area coverage by the measured yield data is below a threshold.

For other cell values estimated by the geostatistical algorithm, the uncertainty index is defined as the percentage of the estimation variance: $100 * (\sigma_j^2/\sigma_{max}^2)$, where σ_j^2 is the estimation variance for cell j and σ_{max}^2 the maximum estimation variance for the whole field. This index accounts for the various factors that influence the estimation error (Isaaks and Srivastava, 1989). The estimation variance depends on the covariance model of the yield variable.

Example data set

A 53.9 ha (133 acre) field, equipped with a center pivot irrigation system, was selected for study in 1996. The field, located in eastern Washington, was variable in topography, soil texture, and soil test NO_3-N, P, and K. Average annual rainfall is less than 250 mm. A short-season potato (*Solanum tuberosum* L.) variety, "Shepody," was planted in the field on March 20–22, 1996. Water and nitrogen were applied through the irrigation system in accordance with standard grower practice. Additional information about the study can be found in Schneider *et al*. (1997).

Potato yield mapping was conducted from late July to early August. HarvestMaster HM-500[1] yield monitors were installed on two four-row Lockwood potato diggers. Yield data were recorded at a 3-s time interval. Sub-meter accuracy Differential Global Positioning System (DGPS) receivers were used to record positions. A time delay based on the total travel distance of the potato mass and belt speed was applied to each record. The actual weight of the potatoes in each truck was measured and used to adjust the recorded yield data.

Results and discussion

A total of 21,282 yield data points were collected in the field (Figure 2). Data were missing on two large bands (**A** and **B**) because yield monitors were not installed until after the first day of the harvest. Several small bands with no yield data occurred when, due to mechanical problems, a yield monitor-equipped digger was replaced with an unequipped digger. Band **C** was the result of alternative swaths situation where only one of the two harvesters was equipped with the yield monitor. Some data collected on the edges of the field were not reliable and were removed from further analysis. The descriptive statistics for the point yield data set (Set A) are
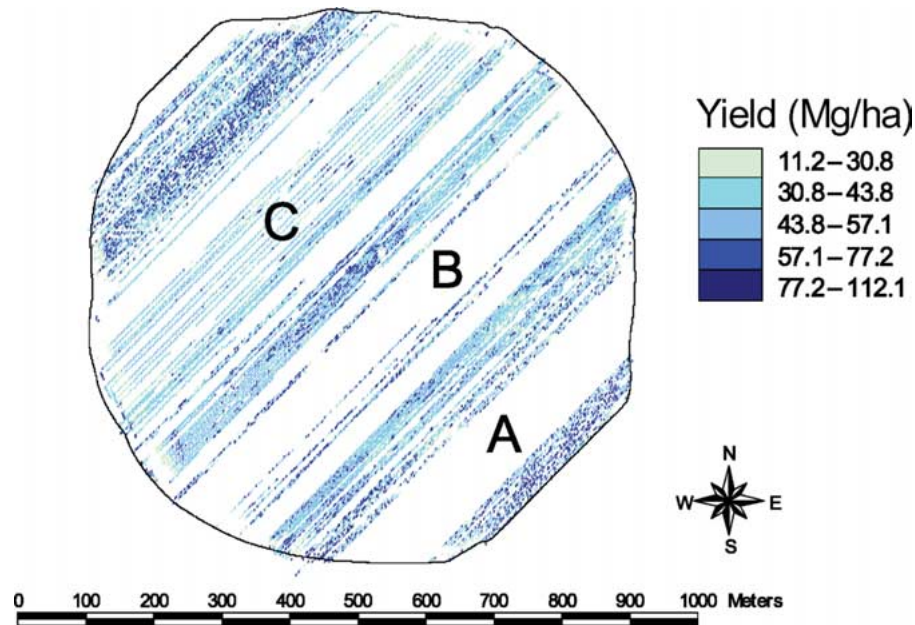
Figure 2.   Yield data collected on the potato field, a total of 21,282 yield points.

Table 1.   Descriptive statistics for different yield data sets

Statistics	Set A[a]	Set B[b]	Set C[c]	Set D[d]	Set E[e]
Number of points/cells	21282	7267	5159	247	241
Mean (Mg/ha)	46.39	43.55	42.66	40.96	42.84
Standard deviation (Mg/ha)	16.47	10.62	8.57	15.11	7.01
Maximum (Mg/ha)	112.09	93.25	90.56	89.44	82.94
Minimum (Mg/ha)	11.23	11.88	13.90	11.88	20.18
Coefficient of variation (%)	35.49	24.38	20.09	36.88	16.37

[a]The original point yield data.
[b]Cells in which at least 20% of the areas are covered by the measured yield points. Cell values are estimated by the moving average method.
[c]Cells in which less than 20% of the area are covered by the measured yield points. Cell values are estimated by the kriging method.
[d]Cells in which 15–20% of the area are covered by the measured yield points. Cell values are estimated by the moving average method.
[e]Cells in which 15–20% of the area are covered by the measured yield points. Cell values are estimated by the kriging method.

given in Table 1. Large spatial variation of the yield, with a 35% coefficient of variation, existed in the field.

A 0.3 m (1 ft) fixed-distance interval d was used to convert the original yield data set into a fixed-distance interval data set, as described in the data conversion section. Cell size for the raster yield map was selected as 6.1 m (20 ft). The moving average
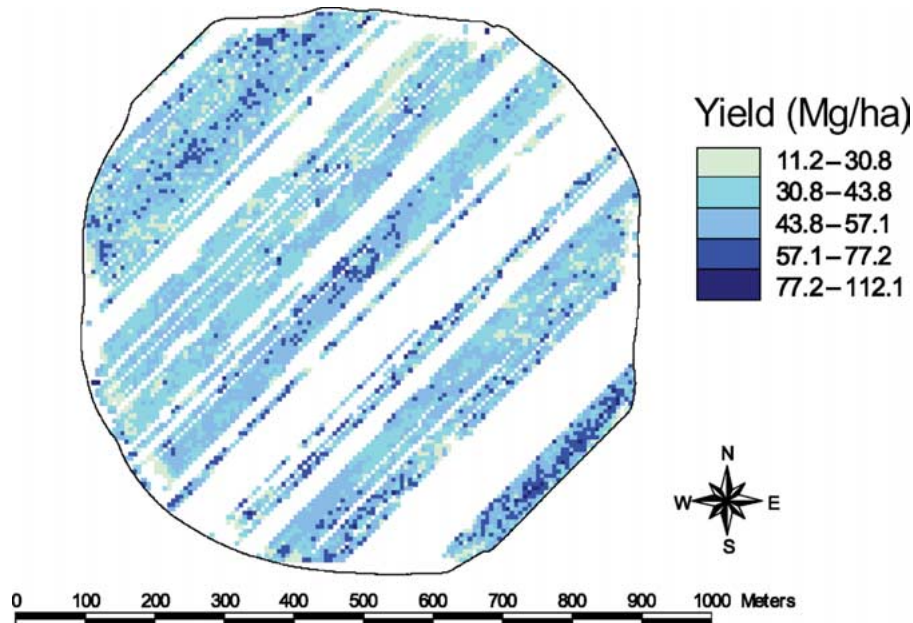
Figure 3.   The partial yield map created by applying the moving average algorithm to those cells in which at least 20% of the areas are covered by the measured yield points.

algorithm was applied to estimate the average yield in a cell when at least 20% of its area was covered by the measured yield points. This resulted in a yield map which shows the yield distribution for 50.7% of the entire field (Figure 3). The descriptive statistics for this yield data set (Set B) are given in Table 1.

For the rest of the cells, the kriging algorithm was applied to estimate their values. As discussed before, the available cell values were used instead of using the original yield measurements in kriging. Spherical semi-variogram models were fitted to the experimental semi-variances. The size of the neighborhood and the maximum number of known values in the kriging equations were chosen as 15.2 m (50 ft) and 20 respectively. The kriging process resulted in a partial yield map which covers an additional 36% of the field. The descriptive statistics for this yield data set (Set C) are also given in Table 1. A combined yield map is shown in Figure 4. Large spatial variability of the potato yield across the field was observed. Spatial correlation among potato yield and quality, field topographical features, soil test nutrient levels, and other factors are reported elsewhere (Schneider *et al.*, 1997).

In order to compare the performance of the moving average method with that of the kriging method, cell values were estimated by both methods for those cells in which 15–20% of the area were covered by the measured yield points. The descriptive statistics for both yield data sets (Set D, E) are given in Table 1. The mean, maximum, and minimum cell values by these two methods were similar. However, the standard deviation and coefficient of variation were much smaller for Set E. This result may indicate that, when the number of measurement points is small within a cell, the cell value is better estimated by the kriging method. On the other hand, when
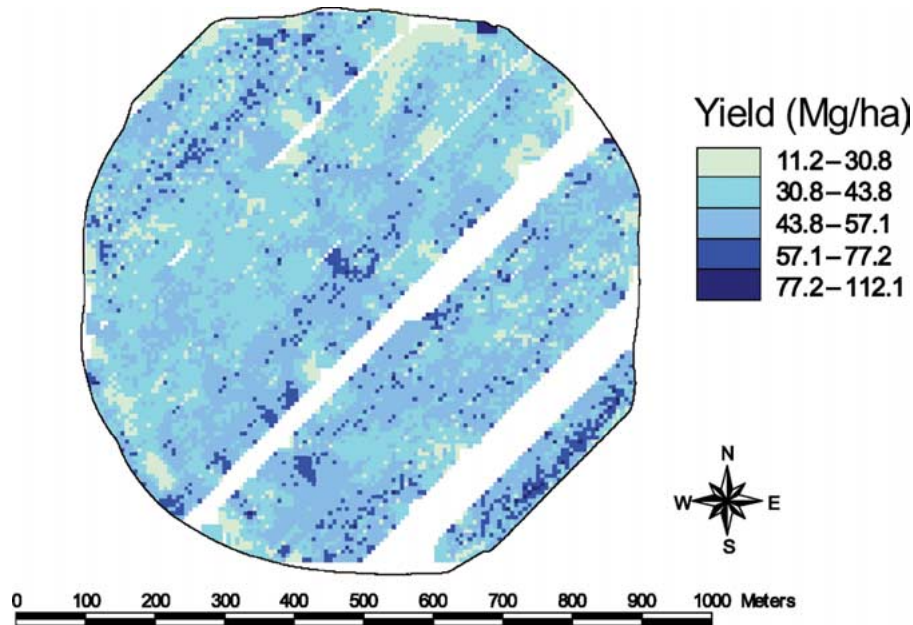
Figure 4. The combined yield map created by applying both the moving average algorithm and the kriging algorithm.

the number of measurement points is large within a cell, the moving average algorithm may provide better estimates.

Uncertainty index maps based on the area coverage of the measured yield points (Figure 5) and the kriging variance (Figure 6) were developed. Clearly, the uncertainty index based on the area coverage (Figure 5) was inversely proportional to the density of the measured yield points (Figure 2). Although this index is easy to calculate, caution needs to be taken with interpretation. Since it only reflects the number of measurements taken to determine the average yield within each cell, the higher uncertainty indexes for some cells may be inflated when the yield distribution is quite uniform. In other words, high accuracy in estimation can be achieved for a uniformly distributed variable with only a limited number of measurements. Practically, crop yield exhibits large spatial variability within a field, and the use of this type of uncertainty index makes a reasonable simplification.

On the other hand, the uncertainty index based on kriging variance has a good theoretical foundation. A cell estimate with smaller kriging variance is always better than an estimate with larger kriging variance. However, the kriging variance is largely dependent on the accurate determination of the yield semi-variogram.

The uncertainty index maps can help make visual judgment on the quality of the related yield map. This index can also be further explored and may be incorporated in some statistical or stochastical models which require yield input along with its range or distribution. Further work is needed to combine the uncertainty index maps generated by both algorithms into one unique uncertainty index map.
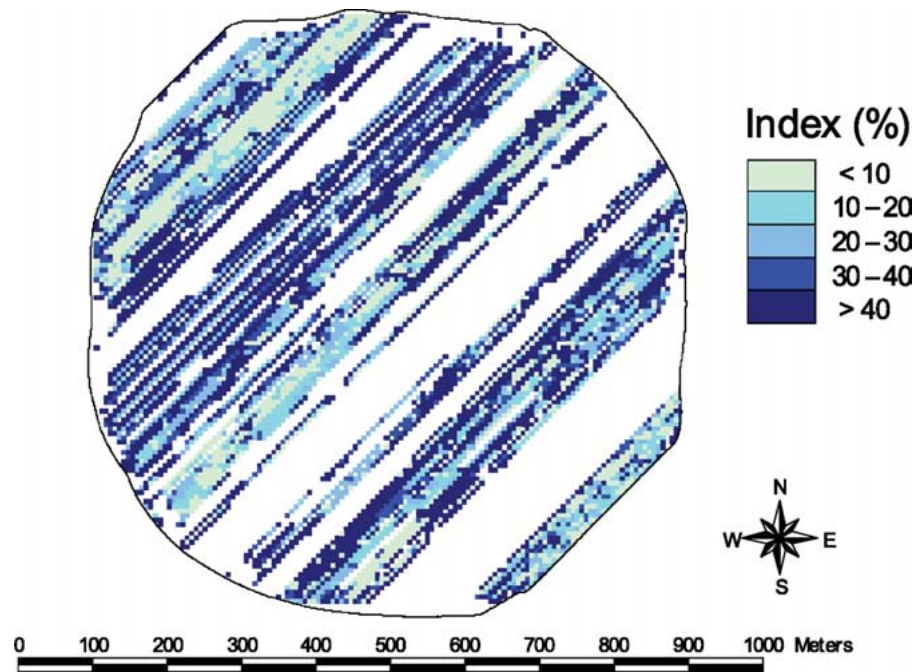
Figure 5.   The uncertainty index map based on the areal coverage of the measured yield points.
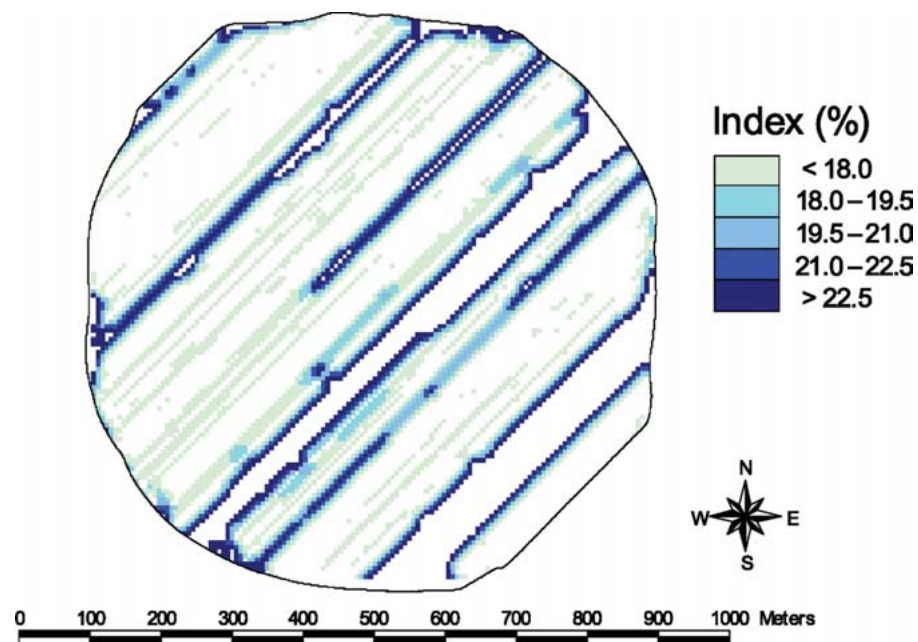

Figure 6.   The uncertainty index map based on the kriging variance.

Conclusions

A procedure for converting the point yield measurement data set into yield maps and for assessing the accuracy of the resulting maps was developed. Two important concepts, data discretization and uncertainty index map, were proposed. The discrete data set in a fixed-distance interval is a realistic approximation of the continuous yield variable. The uncertainty index map allows us to evaluate the reliability and accuracy of the yield maps. The proposed procedure is particularly useful in situations when the measured yield points can not cover the entire field.

Notes

1. Mention of trade names does not imply preferential treatment or endorsement by Deere, USDA, or WSU over similar products available from other sources.

References

Blackmore, B. S. and Marchall, C. J. 1996. Yield mapping: Errors and algorithm. In: *Proceedings of the 3rd International Conference on Precision Agriculture*, edited by P. C. Robert, R. H. Rust and W. E. Larson (ASA, CSSA, SSSA, Madison, WI, USA), pp. 403–415.

Bogaert, P., Mahau, P. and Beckers, F. 1995. The spatial interpolation of agro-climatic data. FAO Agrometeorology Series Working Paper, Number 12 (FAO, Rome, Italy).

Borgelt, S. C. and Sudduth, K. A. 1992. Grain flow monitoring for in-field yield mapping. ASAE Paper No. 92-1022 (ASAE, St. Joseph, MI, USA).

Burgess, T. M. and Webster, R. 1980a. Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging. Journal of Soil Science **31**, 315–331.

Burgess, T. M. and Webster, R. 1980b. Optimal interpolation and isarithmic mapping of soil properties. II. Block kriging. Journal of Soil Science **31**, 333–341.

Han, S., Schneider, S. M., Rawlins, S. L. and Evans, R. G. 1997. A bitmap method for determining effective combine cut width in yield mapping. Transactions of the ASAE **40**(2), 485–490.

Isaaks, E. H. and Srivastava, R. M. 1989. *An Introduction to Applied Geostatistics* (Oxford University Press, New York, USA).

Journel, A. G. and Huijbregts, C. J. 1978. *Mining Geostatistics* (Academic Press, New York, USA).

Matheron, G. 1963. Principles of Geostatistics. Economic Geology **58**, 1246–1266.

Meirvenne, M. V. and Hofman, G. 1989. Spatial variability of soil nitrate nitrogen after potatoes and its change during winter. Plant and Soil **120**, 103–110.

Paz, A., Taboada, M. T., and Gomez, M. J. 1996. Spatial variability in topsoil micronutrients in a one-hectare cropland plot. Communications in Soil Science and Plant Analysis **27**, 479–503.

Reitz, P. and Kutzbach, H. D. 1996. Investigation on a particular yield mapping system for combine harvesters. Computers and Electronics in Agriculture **14**, 137–150.

Schneider, S. M., Boydston, R. A., Han, S., Evans, R. G., and Campbell, R. H. 1997. Mapping of potato yield and quality. In: *Proceedings of the 1st European Conference on Precision Agriculture*, edited by J. V. Stafford (BIOS Scientific Publishers, Oxford, UK), pp. 253–261.

Schneider, S. M., Han, S., Campbell, R. H., Evans, R. G., and Rawlins, S. L. 1996. Precision agriculture for potatoes in the Pacific Northwest. In: *Proceedings of the 3rd International Conference on Precision Agriculture*, edited by P. C. Robert, R. H. Rust and W. E. Larson (ASA, CSSA, SSSA, Madison, WI), pp. 443–452.

Searcy, S. W., Schueller, J. K., Bae, Y. H., Borgelt, S. C. and Stout, B. A. 1989. Mapping of spatially variable yield during grain combining. Transactions of the ASAE **32**(3), 826–829.

Vansichen, R. and de Baerdemaeker, J. 1991. Continuous wheat yield measurement on a combine. In: *Automated Agriculture for the 21st Century: Proceedings of the 1991 Symposium* (ASAE St. Joseph, MI, USA), pp. 346–355.

Walter, J. D., Hofman, V. L. and Backer, L. F. 1996. Site-specific sugarbeet yield monitoring. In: *Proceedings of the 3rd International Conference on Precision Agriculture*, edited by P. C. Robert, R. H. Rust and W. E. Larson (ASA, CSSA, SSSA, Madison, WI, USA), pp. 835–844.